Linked Open Census Data

Merono, A.; Ashkpour, A.; Scharnhorst, A.; Gueret, C.D.M.; Wyatt, S.

published in
DHCommons
2015

document version
Publisher's PDF, also known as Version of record

CEDAR: Linked Open Census Data

From fragment to fabric. Dutch census data in a web of global cultural and historic information

Albert Meroño-Peñuela, Ashkan Ashkpour, Andrea Scharnhorst, Christophe Guéret and Sally Wyatt

Introduction

Census Data Open Linked. From fragment to fabric - Dutch census data in a web of global cultural and historic information[1] (CEDAR) is an ongoing (2011-2015) Dutch multidisciplinary national research project. It is funded by the Royal Netherlands Academy of Arts and Sciences[2] (KNAW) as part of the Computational Humanities Programme[3]. Its participants are Data Archiving and Networked Services[4] (DANS), the VU University Amsterdam[5], the International Institute of Social History[6] (IISH) and the Erasmus University Rotterdam[7].

The overall goal of the project is to provide an easier access to the Dutch historical census data. The intended research audience of the project are historians and other humanities scholars interested in historical statistical information. To understand the importance of CEDAR one has to know that since decades efforts have been made by the Central Bureau voor de Statistiek[8] (CBS) (Dutch Central Statistical Office), DANS and others to make the Dutch Historic Census better available for the wider public as well as for research. In the

[1] See http://www.cedar-project.nl/
[2] See http://knaw.nl/
[3] See http://ehumanities.nl/
[4] See http://dans.knaw.nl/
[5] See http://vu.nl/
[6] See http://socialhistory.org/
[7] See http://www.eur.nl/
[8] See http://www.cbs.nl

Netherlands we have sources from census data going back to 1795. Up to 1971 in each decade a census has been carried out, with different questions and different also in granularity of collected information. The primary remaining sources are books in which tables have been published containing the aggregation of census information. Those books have been scanned. Later, a data entry project has been carried out to transfer the tables into Excel files. Both images and excel files have been partially indexed and made available via a Content-Management-System for browsing and some search capabilities. However, the digital representation of the Dutch Historic Census in this form is not machine readable, and concerning current Big Data efforts in the Humanities quite outdated. This was the motivation to set up CEDAR, and those Excel files are the heritage from which this project started[9].

CEDAR seeks to answer fundamental questions about social history in the Netherlands and the world in automatic, web-scalable and reproducible ways. More concretely, the aim of CEDAR is to publish the Dutch historical censuses (1795-1971) in the Semantic Web, using this dataset as a starting point to build a semantic data-web of socio-historical information. With such a web we will be able to more easily answer questions such as:

- What kind of patterns can we identify and interpret in expressions of regional identity?
- How to relate patterns of changes in skills and labour to technological progress and patterns of geographical migration?
- How to trace changes of local and national policies in the structure of communities and individual lives?

Sometimes, census data alone are not sufficient to answer these questions. CEDAR exploits Web standards[10] to make census data interlinkable with other hubs of historical socioeconomic and demographic data. When integrated, these hubs can better support the historical research cycle. The project will result in generic methods and tools to weave historical and socio-economic datasets into an interlinked semantic data-web.

This broad aim touches unavoidably upon many interdisciplinary research areas and audiences. Publishing socio-historical data on the Web in a semantically rich and consistent manner poses fundamental challenges for Knowledge Representation and Reasoning, two of the key fields in Artificial Intelligence (AI). The deployment of tools and methods to achieve these goals in a reproducible and efficient way is closely related with Software Engineering and Computing. On the other hand, Social History, located at the crossroads between history and social sciences, produces fundamental research questions about social change and suggests domain-specific models and standards[11] for socio-historical data. The interplay within Computing and the Humanities (the basic components of the Digital Humanities) in CEDAR works two-ways: (a) we use AI and Computing to give infrastructure, scale, formalism

[9] See legacy data at http://www.volkstellingen.nl and https://github.com/CEDAR-project/DataDump/tree/master/xls
[10] See http://www.w3.org/standards/
[11] See http://www.clio-infra.eu

and reproducibility to address Social History issues; and (b) we use Social History to inspire AI and Computing with new algorithms, methods and tools.

Related Work

Starting in 1996, major efforts have been undertaken in the digitization of the Dutch historical censuses. In order to provide better access to censuses, the Dutch Statistics (CBS) and the Data Archiving and Networked Services (DANS) institute, grown out from NIWI (Het Nederlands Instituut voor Wetenschappelijke Informatiediensten) cooperated in the digitization of the aggregated results from the censuses of 1795-1947. Another goal of this cooperation was also to improve the accessibility of the 1960 and 1970 censuses which cover so-called micro data (i.e. registers with precise information about each individual). The first step in the digitization process of the Dutch historical censuses focused more on 'medium conversion', resulting in thousands of scans of the books which were published as images. The second major activity was the 'manual' conversion (data entry) of these images into Excel files, resulting in over 2300 disconnected and heterogeneous excel files with different levels of granularity.

CEDAR also strongly builds on the outcomes of the *Hub for Aggregated Social History* (HASH) project (2007-2011, Netherlands Interdisciplinary Demographic Institute, NIDI-KNAW, and Radboud University Nijmegen, RU). HASH aimed to realize a web hub for accessing relevant demographic, social, economic and political data on all Dutch historical municipalities (1812-2000). This was done in four ways: 1) by complementing and correcting datasets that already exist at several institutions, 2) by synchronizing these datasets by means of standardized meta-information, 3) by creating a web portal for immediate access to the data, and 4) by creating an innovative interface for selecting and visualizing the data. Thanks to this workflow, the metadata of the Dutch historical census tables are indexed and can be queried. CEDAR seeks to extend this with fine-grained access not only to metadata, but also to aligned, harmonized and ready-to-process census data, exposed on the Web in a machine-readable and processable way.

First introduced in 2001 by Berners-Lee, Hendler and Lassila [2], the Semantic Web was conceived as an evolution of the existing Web (based on the paradigm of the document) into a Semantic Web (based on the paradigm of meaning and structured data). Concretely, the Semantic Web can be defined as the collaboration and the set of standards that pursue the realization of this vision. The World Wide Web Consortium[12] (W3C) maintains the Resource Description Framework (RDF), the basic layer on which the Semantic Web is built. RDF is a set of W3C specifications designed as a metadata data model. It is used as a conceptual description method: entities of the world are represented with nodes (e.g. *Dante Alighieri* or *The Divine Comedy*), while the relationships between these nodes are represented through

[12] See http://www.w3.org/

edges that connect them (e.g. Dante Alighieri *wrote* The Divine Comedy). Linked Data[13] is the method of publishing and interlinking structured data on the Web using RDF and standard vocabularies[14]. The Linked Open Data cloud[15] is the set of all linked datasets on the Web that use Linked Data.

There have been multiple efforts to use semantic technologies and Linked Data approaches to represent, publish and interlink historical data, including historical censuses, meaningfully on the Web. Meroño-Peñuela et al. [2] offer an up-to-date survey on this subject. More closely related approaches on publishing census data also exist. The 2000 U.S Census[16] has an RDF version which provides population statistics on various geographic levels, although the dataset is not historical and it does not harmonize different time-gapped releases. The Canadian health census uses LOD principles to provide greater access to the data and to promote greater interoperability, unachievable with conventional data formats [3]. Similarly, in the context of a national large scale project regarding the management of socio-demographic data in Greece, Petrou et al. [4] have applied LOD techniques to the Greek population census of 2011. A similar goal here is to publish 'traditional' datasets into RDF and allow easier access and use of the census by e.g. third parties, aiming to develop a platform within which the Greek census is converted, interlinked and available in a LOD format. The 2001 Spanish Census project is another advocate of applying LOD principles to census data, while encouraging the development of open government initiatives [5]. Using microdata from the 2001 population census, the authors of the Spanish Census project propose a solution for converting the data into open formats allowing greater discoverability, accessibility and integration; a recurrent topic in all of the mentioned projects.

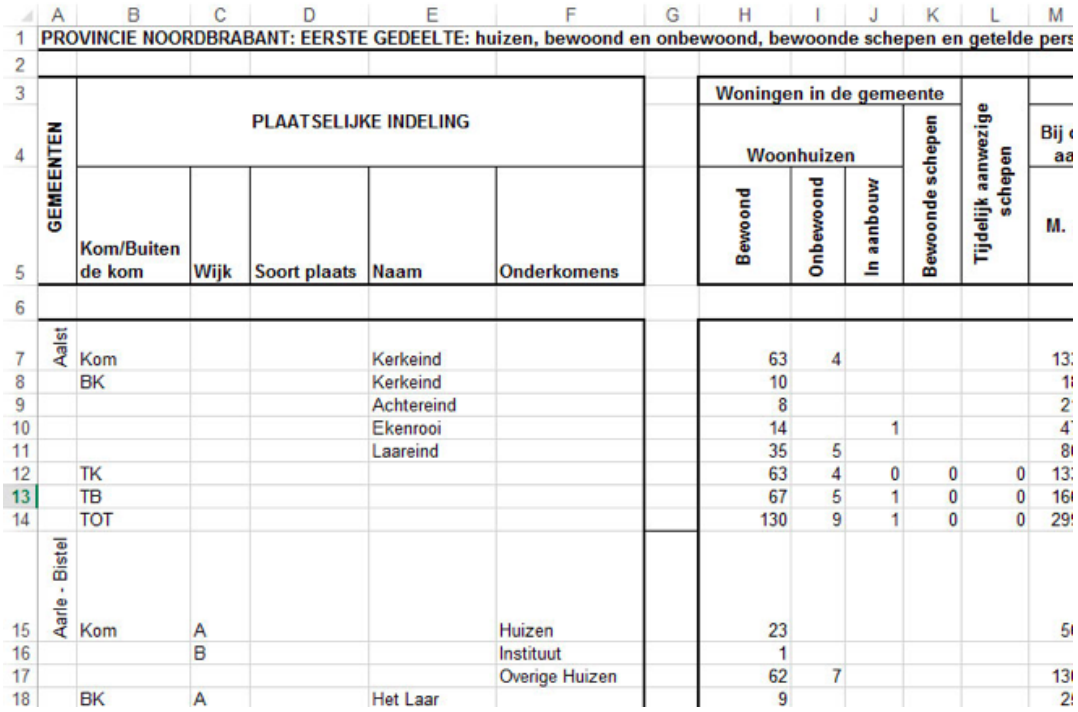
All of the above projects have harmonized RDF census data within the domain of each census year, using micro data as a starting point. Therefore, a hypothetical methodology to harmonize different time-gapped versions of aggregated Linked Census Data, especially by leveraging the externally linked datasets, remains an open, unsolved research problem.

Methodology

The primary format of the Dutch historical censuses (1795-1971) are images of scanned books. A small subset of the 300,000 images[17] is available as 507 Excel workbooks, containing 2,288 tables (0.7 % of the data) (see Figure 1). Meaningful historical information is currently hidden in these tables. In order to fully reap the benefits of this dataset, temporal comparisons and cross-connections between these tables are required. However, these

[13] See also http://www.w3.org/DesignIssues/LinkedData.html
[14] A summary of these vocabularies can be found at http://lov.okfn.org/dataset/lov/
[15] A summary of contained datasets can be found at http://linkeddata.org/
[16] See http://datahub.io/dataset/2000-us-census-rdf
[17] The biggest subset of it is currently available at https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:44159

tables are highly dispersed, hardly comparable, differently aggregated and non-trivially queryable, mainly due to the temporal gap (about 10 years) between the versions, and the lack of data harmonization.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	PROVINCIE NOORDBRABANT: EERSTE GEDEELTE: huizen, bewoond en onbewoond, bewoonde schepen en getelde pers												
2													
3								Woningen in de gemeente					
4	GEMEENTEN	PLAATSELIJKE INDELING						Woonhuizen			Bewoonde schepen	Tijdelijk aanwezige schepen	Bij aa
5		Kom/Buiten de kom	Wijk	Soort plaats	Naam	Onderkomens		Bewoond	Onbewoond	In aanbouw			M.
6													
7	Aalst	Kom			Kerkeind			63	4				13
8		BK			Kerkeind			10					1
9					Achtereind			8					2
10					Ekenrooi			14		1			4
11					Laareind			35	5				8
12		TK						63	4	0	0	0	13
13		TB						67	5	1	0	0	16
14		TOT						130	9	1	0	0	29
15	Aarle - Bistel	Kom	A			Huizen		23					5
16			B			Instituut		1					
17						Overige Huizen		62	7				13
18		BK	A		Het Laar			9					2

Figure 1. Example of one of the Excel workbooks with a census table.

In order to solve these issues, we use a two-fold approach that combines *Linked Open Data (LOD) principles* with *harmonization practices*. On the one hand, we represent the census dataset as LOD using Web standards, making it interlinkable with other hubs of historical socioeconomic and demographic information. On the other hand, we apply state-of-the-art harmonization techniques [6,7,8,9] to clean, normalize and make the data compatible and comparable.

Linked Census Data

Following standard LOD guidelines[18], we exploit the Resource Description Framework[19] (RDF), the W3C Web standard for data publishing and exchange on the Web. We transform the 507 Excel workbooks of the dataset into RDF and following the Linked Data paradigm, creating historical Dutch Linked Census Data for the first time. Making extensive use of semantic technologies and LOD principles, we link this dataset to other relevant datasets using standard vocabularies, leveraging specific data classifications, taxonomies and ontologies.

[18] See http://linkeddatabook.com/editions/1.0/
[19] See http://www.w3.org/standards/techs/rdf

First, we select a standard data model and a set of standard vocabularies that fit the data. Our choice for the former is the RDF Data Cube vocabulary[20] (QB), the W3C standard for publishing multi-dimensional data, such as statistics, on the Web in such a way that they can be linked to related datasets and concepts. QB allows us to express the same observations contained in the Excel spreadsheets in RDF, without losing meaning. We extend the model to also preserve the original layout.

Besides QB, the converted dataset in RDF makes use of the following additional vocabularies:

- DCMI Metadata terms: http://purl.org/dc/terms/
- Web Ontology Language (OWL): http://www.w3.org/2002/07/owl#
- RDF Schema: http://www.w3.org/2000/01/rdf-schema#
- Simple Knowledge Organization System (SKOS) RDF Schema: http://www.w3.org/2004/02/skos/core#
- XML Schema Definition (XSD): http://www.w3.org/2001/XMLSchema#

Once transformed into RDF, we refer to the resulting dataset as CEDAR's *raw data* layer. The RDF graph representation of one cell is shown in Figure 2.

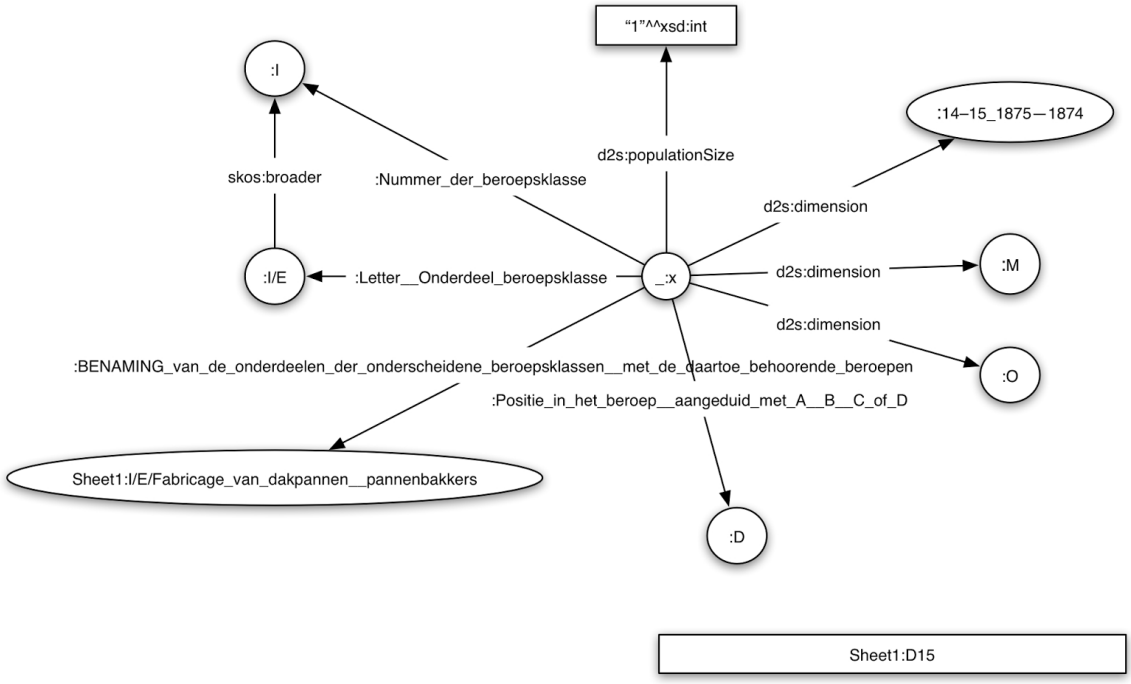


Figure 2. RDF graph representation of one of the cells in the census tables.

[20] See http://www.w3.org/TR/vocab-data-cube/

Census Harmonization

Although RDF allows us to represent the data in the tables in a fine granular way, the problem of harmonizing the data remains unsolved. In order to implement this harmonization, we apply a two tier model (see Figure 3): an upper tier, allowing access to harmonized census data; and a lower tier, allowing access to the original historical primary sources. We implement these two tiers in three modules: the *raw data* (containing a direct translation of the numbers and concepts in the tables, as described before), the *annotations* (with corrections and comments from the dataset creators and curators), and the *harmonization* (linking and standardizing heterogeneous census entities).

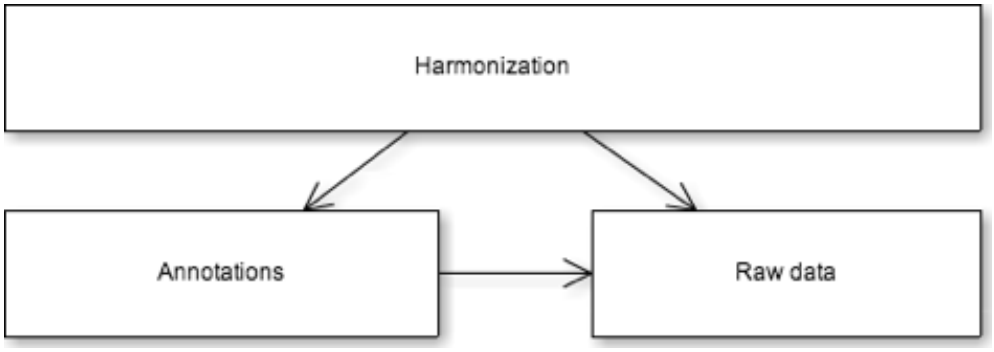


Figure 3. Harmonization, annotation and raw data layers.

When dealing with historical census data researchers need to make sense of all the irregularities in structures, classifications and deal with redundant, inconsistent or erroneous data. Currently harmonization is still very much a loose term for researchers, encompassing different views and methods on how data should be restructured. When working with historical censuses, due to their nature, researchers are often forced to create their own classifications or use existing systems for comparative research. In historical research harmonization is often defined as the creation of a unified, consistent data series from disparate census samples when dealing with historical census data [6]. As harmonization is not a standard process we want to leave room for experimentation and different interpretations of the data. Harmonization therefore strongly relies on interpretation and specific goals. Moreover, access to source data must always be guaranteed. We separate the *raw data*, *annotations* and *harmonization* layers to accommodate this (see Figure 2). Harmonization of aggregated statistical data consists of a set of practices such as rules, data transformations, cleaning, standardization, use of standard classification systems, data smoothing, interpolating, extrapolating etc. We apply all these techniques in the *harmonization layer*, building on top of data in the *raw layer* and always considering the expert annotations in the *annotations layer* (see Figure 2).

Contribution

One of the earlier tasks of CEDAR was to convert the original census tables contained in Excel tables into RDF. Unfortunately, none of the tabular-to-RDF conversion tools available[21] was suitable for CEDAR: heterogeneity of the layout, column and row headers spanning various cells, and the lack of semantic expressivity made clear that a special-purpose tool was needed. To solve this, we coded, together with Data2Semantics[22] (COMMIT, VU University Amsterdam), TabLinker[23], a supervised Excel-to-RDF-QB converter. TabLinker reads markup of Excel files to produce faithful RDF QB representations, works virtually with any Excel file in a generic way, and can be customized by several parameters.

Using TabLinker and an expert-based markup of the source files, we transformed the Dutch historical census tables into RDF. Loaded into a triplestore, the entire graph database is available for users and machines alike to query live on the Web via a SPARQL (a SQL-like standard language to query Linked Open Data) endpoint[24]. There is work in progress on generating SPARQL documentation to enable historians, social scientists and humanities scholars (in addition to computer scientists) to write their queries. Complete dumps of the converted data are also available for download[25].

However, the conversion alone does not produce an harmonized dataset: many values differently spelled need to be mapped together, variables need to be aggregated at different geographical levels, etc. To address harmonization, we take several approaches:

- We first harmonize automatically as many variables as we can, with a set of harmonization scripts that produce human-readable documentation[26]
- We then ingest these automatic harmonization into CEDAR Harmonize[27], a web interface that allows knowledge experts to fine-tune the harmonization process
- In order to standardize common statistical historical variables, like demographic structures, housing types, occupational classes and statuses, or religious denominations, we build bottom-up classification systems from the raw data. To achieve this, we have developed TabCluster[28] [10], an algorithm that builds taxonomies out of flat lists of values by combining lexical attributes (using hierarchical clustering) and enriches them semantically (using knowledge in DBpedia and Wordnet).

---

[21] See http://www.w3.org/wiki/ConverterToRdf#Excel
[22] See http://www.data2semantics.org/
[23] https://github.com/Data2Semantics/TabLinker/
[24] The SPARQL endpoint is available at http://lod.cedar-project.nl:8080/sparql/cedar
[25] Downloads available at https://github.com/CEDAR-project/DataDump
[26] See https://github.com/CEDAR-project/Integrator
[27] See https://github.com/CEDAR-project/Harmonize
[28] See https://github.com/CEDAR-project/TabCluster

In order to produce a 5-star dataset[29] we produce links that connect the CEDAR dataset to other LOD datasets. Concretely, we issue links (see Figure 4)[30]:

- To the Historical International Standard Classification of Occupations[31] (HISCO)
- To the Amsterdamse Code and URIs of gemeentegeschiedenis.nl
- To occupations in the ICONCLASS[32] system
- To/from the Dutch Ships and Sailors[33] dataset

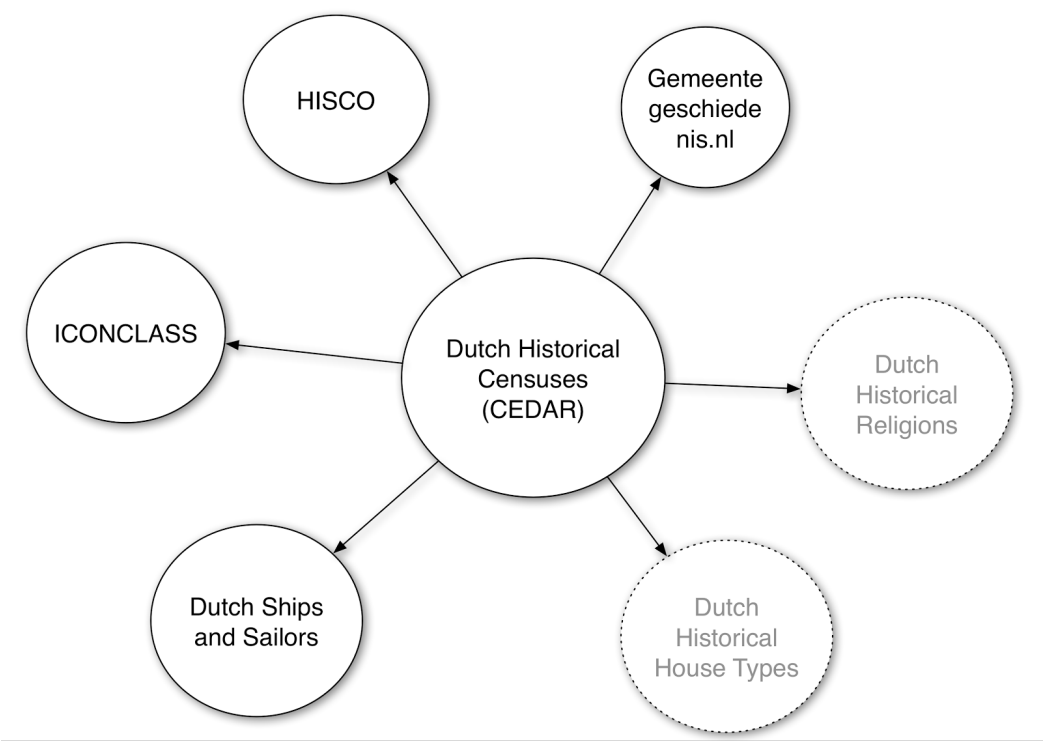


Figure 4. Linked datasets to/from CEDAR.

CEDAR aims at high reproducible research. Ideally, we target a platform that can run the complete transformation and harmonization pipeline, integrating all the necessary steps to produce a top quality LOD dataset. To achieve this, we are developing the CEDAR Integrator [34], a platform integration workflow that will automatise the semantic publication pipeline, from isolated Excel files to RDF ready-to-publish Linked Census Data.

[29] See details at http://www.w3.org/DesignIssues/LinkedData.html
[30] All linksets are available online at https://github.com/CEDAR-project/DataDump/tree/master/links
[31] See http://historyofwork.iisg.nl/
[32] See http://iconclass.org/
[33] See http://ghhpw.com/ships_and_sailors.php
[34] See https://github.com/CEDAR-project/Integrator

A fundamental aspect of CEDAR is to contribute to using Artificial Intelligence and Computing to support the historical research cycle. To this end, we have conducted a survey on the state-of-the-art of applying semantic technologies to historiography [1]. But beyond this, we are concerned about delivering tools and systems to social historians that support historical research. We do this through the following list of concrete contributions.

Comparability: Layout and semantics

The main pitfall of the original form of the dataset, Excel spreadsheets, is that variance and irregularities in *(a)* the layout, and *(b)* the definitions of statistical variables in these tables hampers an easy access to the historical data they contain. As a result, historians need to manually open these files one by one (the dataset contains more than 2,000 of them) and manipulate data items in a non-reproducible way, in order to get data that provides answers to their research questions. With the data model provided by RDF Data Cube (see Methodology section) we overcome irregularities in the interpretation of these changing layouts. With the harmonisation and integration workflows (see Linked Census Data and Census Harmonization sections), we overcome irregularities in varying definitions of statistical variables. Combining the two, we generate, for the first time, a database which is longitudinally accessible and comparable by historians: meaning that they don't have to deal with disparate files, and are able to get answers to their questions using the SPARQL query language.

Standardisation of historical statistical variables

The representation of the Dutch historical censuses as Linked Data makes it very easy to link data points of the dataset to other datasets. This way, for instance, we can say that the value *'vrouwen'* (women) of the population count in Amsterdam in 1889 is the same value '*SDMX:Female*' of the variable '*SDMX:Sex*' of the international standards used to define statistical variables about sex (SDMX[35]). We use this easy linking to standardise values in historical data in general, like historical occupations, historical religions, historical house types, etc. Without such easy linking, historians would need to create classification systems and variables for every specific case. With proper standardisation, we also make it easier for data consumers to better understand and use the dataset.

Enrichment from/to other datasets

Similarly, the representation of the Dutch historical censuses as Linked Data allows us to link resources of the dataset with resources of other Web datasets easily. For instance, users can confront the population counts of Dutch important cities in the 18th, 19th and 20th centuries (from our dataset) with the current population counts available on the Web (from e.g. DBpedia, the Linked Data version of Wikipedia) using a single SPARQL query against both data sources. This way, users of the census data can enrich their queries with other Linked Data they find on the Web that they consider useful or interesting to merge. Likewise, other data sources can be enriched exactly the same way, by querying the census data from our

[35] Statistical Data and Metadata Exchange; see http://sdmx.org/

SPARQL endpoint and adding it to the results.

Automation of longitudinal analysis

Combining the three previous points, historians have now a uniform and automated access to the information contained in the original tables[36]. Moreover, the publishing of these data as Linked Data on the Web makes it possible to anybody to reference any single data point through its URI (and to dereference it to retrieve useful information), contrary to what happens in closed systems, where data (even if integrated, harmonised and standardised) is left in an isolated and non-accessible or referenceable way.

Study of change over time

Last, but not least, such a uniform access opens up the study of social change in the Netherlands in the period 1795-1971. Moreover, we study the fundamental question of how historical concepts have drifted in meaning over time, through the use of statistics and Machine Learning over different archived dataset/ontology versions[37] [11]. Such study was very deeply hampered before, due all issues discussed before. Now, automatic conversion, linkage and standardisation are allowing us to develop visualizations and statistical analyses[38] that allow historians to better understand phenomena beyond the unmanageable original format of the data, at the same time they help on detecting and solving errors in the proposed methodology.

We also enable the exploration and validation of our results by ongoing development of analysis tools and visualizations[39]. This way, we provide methods to leverage Linked Census Data by humanities scholars, including historians, social historians and social scientists, allowing fine-grained querying via SPARQL (experimentally, also via natural language interfaces like *hald*[40]), and linked data discovery and exploitation via visualization and Linked Data browsing[41]. In general, we support the historical research cycle by adding as much automation, intelligence and semantics as we can. Prototypes for web interfaces gathering an integrated access to all browsing features have been developed in a parallel bachelor thesis[42].

We plan preservation and sustainability of CEDAR data and services at different levels. First, we make available all data and source code via a GitHub profile[43], open for any users and developers who wish to collaborate. Second, we strongly collaborate with DANS to archive research results (data and tools) in EASY[44], the Trusted Digital archive of DANS[45]. We are

---

[36] See e.g. http://lod.cedar-project.nl/cedar/data.html
[37] Experiment results and source code available at https://github.com/albertmeronyo/ConceptDrift
[38] See e.g. http://lod.cedar-project.nl/cedar/stats.html
[39] See http://goo.gl/I3xyYz, http://goo.gl/JHEjXL and http://www.cedar-project.nl/visualizing-sparql-query-results-on-the-census/
[40] See repository at https://github.com/CEDAR-project/hald
[41] See http://lod.cedar-project.nl:8888/cedar/ for an interface that allows browsing all published census Linked Data
[42] See http://dutchdatagroup.nl/volkstellingen
[43] See https://github.com/CEDAR-project
[44] See https://easy.dans.knaw.nl/

especially interested in all issues that preservation of Linked Data arises[46].

Most of the described contributions are work in progress. Technical documentation and reports are constantly updated and available online[47].

Peer Review

CEDAR is part of the Computational Humanities Programme of the KNAW, and is part of the eHumanities group[48]. There was a closed call for projects within the humanities institutes of the KNAW. All proposals were sent for international peer review, and final decisions were made by the Computational Humanities Programme Committee, consisting of senior scholars in digital humanities based in Dutch universities. CEDAR is composed of two PhD positions and one postdoc position. The recruitment of staff for those positions took place in an open competition.

CEDAR meets regularly once a week, and organizes an annual international symposium[49] where all developments and publications are shared and discussed. We invite international guests from related research fields and senior scholars from in and out the KNAW, and welcome any interested attendant - we are keen on being open. CEDAR also participates in the annual eHumanities group symposium.

Policy

Principles of data management and curation are core to CEDAR. Once finished, CEDAR will deposit all the data to EASY, a Trusted Digital Repository, which will be given proper persistent identifiers. We will motivate users to properly cite the data, by leveraging these identifiers and the applied semantic representations. Thanks to our LOD approach, the provenance of any census data item, down to its data source, can be traced back following semantically rich web links - something that can only be achieved following LOD methodologies.

CEDAR has been chosen as a use case for a current European project: PRELIDA – Preserving Linked Data[50]. This project aims to prepare guidelines to archive and preserve Linked Data. It addresses questions about where to draw a boundary when preserving part of the Linked Open Data cloud, and how to not only archive the data and data models, but also

---

[45] See https://assessment.datasealofapproval.org/assessment_101/seal/html/
[46] See http://www.prelida.eu/
[47] See https://docs.google.com/document/d/1Y7oPwM155Xz2YBNRb1m2-gAXIQfY47_5Ai0GwhwcVtk/pub and https://docs.google.com/document/d/1iMr65cC4tuSfKvqC6Y3c9tzNtOphXQ7DBj9I3o1811o/pub
[48] See http://www.ehumanities.nl
[49] See http://www.cedar-project.nl/cedar-minisymposium-march-1st-2013/ and http://www.cedar-project.nl/cedar-minisymposium-march-1st-2013/
[50] See http://www.prelida.eu

the software which allows to deploy them [12].

Original census materials, as well as the original CEDAR dataset, are open data and owned by the Central Bureau voor de Statistiek[51] (CBS). All tools developed in CEDAR are open source (specifics on licensing pendant).

References

1. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F. Semantic Technologies for Historical Research: A Survey. *Semantic Web Journal* (to appear) (2014)
2. Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.
3. Ahmad C. Bukhari and Christopher J. O. Baker. The canadian health census as linked open data: towards policy making in public health. In Data Integration in the Life Sciences, 2013.
4. Irene Petrou, George Papastefanatos: Publishing Greek Census Data as Linked Open Data. ERCIM News 2014(96) (2014)
5. Javier D. Fernández, Miguel A. Martínez-Prieto, and Claudio Gutiérrez. Publishing open statistical data: the Spanish census. In Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (dg.o '11). ACM, New York, NY, USA, 20-25 (2011)
6. A. Esteve, M. Sobek. Challenges and Methods of International Census Harmonization. *Historical Methods*, 36(2), 37-41 (2003)
7. S. Ruggles. The Minnesota Historical Census Projects. Historical Methods, 28(1), 6-10 (1995)
8. Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., Schlobach, S. Linked Humanities Data: The Next Frontier? A Use-Case in Historical Census Data. proceedings of the 2nd International Workshop on Linked Science 2012 (LISC2012), ISWC 2012, Boston, USA (2012)
9. Lee Emma Williamson. Methods for Coding c19th and c20th Cause of Death Descriptions from Historical Registers to Standard Classifications. European Social Science History Conference 2014, Vienna (2014) https://esshc.socialhistory.org/esshc-user/program/?day=15&time=32&paper=2709&network=36
10. Albert Meroño-Peñuela, Ashkan Ashkpour, Christophe Guéret. From Flat Lists to Taxonomies: Bottom-up Concept Scheme Generation in Linked Statistical Data. Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014), ISWC 2014, Riva del Garda, Italy (2014).
11. Albert Meroño-Peñuela, Christophe Guéret, Rinke Hoekstra, Stefan Schlobach. Detecting and Reporting Extensional Concept Drift in Statistical Linked Data.

[51] See http://www.cbs.nl/

Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC 2013, Sydney, Australia.

12. Sotiris Batsakis, David Giaretta, Christophe Gueret, Rene van Horik, Maarten Hogerwerf, Antoine Isaac, Carlo Meghini, Andrea Scharnhorst (2014) PRELIDA Deliverable 3.1. - State of the art assessment on Linked Data and Digital Preservation. (With further comments and text contributions from Albert Meroño-Peñuela, Peter Doorn, Marat Charlaganov and Menzo Windhower). Project report. Web resource http://www.prelida.eu/sites/default/files/D3.1%20State%20of%20the%20art.pdf